개념 계층을 이용한 스트리밍 데이터의 관리 기법[1)]

한창희°, 박 석
서강대학교 컴퓨터학과
{han°, spark}@dblab.sogang.ac.kr

Streaming Data Management Technique using Concept Hierarchy

Chang-Hee Han°, Seog Park
Dept. of Computer Science, Sogang University

요 약

센서 네트웍, 유비쿼터스 컴퓨팅 환경으로 발전하면서 스트리밍 데이터와 같이 무한한 데이터의 처리에 대한 요구가 많이 커지고 있다. 스트리밍 데이터에 대한 질의 처리는 크게 실시간으로 처리가 요구되는 질의와 과거 데이터에 대한 동향, 근사치 요청질의로 나누어질 수 있다. 기존의 스트리밍 데이터 처리에 대한 연구들은 실시간 질의 처리만을 고려하고 과거 데이터에 대한 질의에 대한 고려는 미약하다. 그리고 사용자가 과거의 데이터에 대한 동향 분석을 요청하는 질의, 또는 과거 어느 시점의 데이터에 대한 요청 혹은 근사치를 요구하는 질의에 대해서는 처리를 할 수 없는 한계점이 있다. 본 논문에서는 스트리밍 데이터 프로세서의 메모리의 범위를 넘어서서 삭제되는 과거 데이터를 디스크의 I/O처리 속도에 맞추기 위해서 로드 셰딩 기법을 적용해서 저장한 후에 개념 계층을 이용해서 사용자가 원하는 데이터만을 효과적으로 저장하는 기법을 제안한다.

1. 서 론

유비쿼터스 컴퓨팅(ubiquitous computing)이 대두하면서 점차 모든 기기들 간의 교환되는 데이터(data)가 점차 많아지면서 스트리밍(streaming) 데이터에 대한 처리가 중요시 되고 있다. 예를 들어 센서 네트웍(sensor network)에서의 유선으로 연결된 센서(이동 기기와 달리 고정 전원을 공급받는 센서)로부터 실시간(real time)으로 전송되는 정보는 연속적으로 무한하게 전송된다. 이러한 스트리밍 형태의 데이터에 대해서 발생할 수 있는 질의는 실시간 질의, 연속적인 질의, 과거 데이타에 대한 질의로 분류할 수 있다. 지금까지의 연구는실시간 질의와 연속적인 질의에 대해서 많이 진행되어왔다. 그렇지만 스트리밍 데이터는 연속적이고 무한한 데이터이기 때문에 한번만 읽을 수 있는 특성으로 인해 제한된 메모리 크기 만큼의 데이터에 대해서만 질의가 가능했다. 다시 말해서 질의의 대상이 되는 데이터의 범위가 메모리에 저장된 데이터보다 이전의 데이터에 대한 처리는 고려하지 않고 있다.

[그림 1]은 시간에 따른 질의의 변화를 나타내고 있다. 사용자가 과거의 시점 C를 기준으로 과거 1년 동안의 데이터의 변화 동향 질의를 요청하고 시스템의 저장 공간이 1개월 분의 스트리밍 데이터만을 저장할 수 있다면 기존의 시스템들은 제한된 저장 공간으로 인해서 1개월 이전의 데이터를 저장하고 있지 않기 때문에 사용자의 질의를 처리할 수 없는 문제점이 발생한다.

[그림 1] 시간에 따른 질의

이와 같은 문제점을 해결하기 위해서 히스토그램(histogram)이나 통계정보(statistic Information)와 같이 RDB에서 연구된 기법들을 적용한다면 시간을 중심으로 하는 질의를 처리가 불가능하고 고정된 데이터에 대한 통계정보를 생성하는 방식을 스트리밍 데이터와 같이 연속적으로 변화하는 데이터에 적용한다면 통계정보 생성의 과부하 문제가 발생하게 된다. 그렇기 때문에 모든 스트리밍 데이터를 저장하지 않고 로드셰딩(load shedding)과 같이 시스템이 처리할 수 있는 한도 내에서 데이터를 저장하고 저장된 데이터에 대해서는 개념 계층(concept hierarchy)을 이용해서 관리한다면 동향이나 근사치 요청 질의에 효과적이다. 또한 사용자 개개인의 요구대로 과거 데이터의 동향을 효과적으로 보여줄 수 있을 것이다.

본 논문은 스트리밍 데이터에 대한 질의 처리에서 과거 데이터에 대한 동향이나 근사치 요청 질의 처리를 가능하게 하기 위해서 개념 계층을 이용한 기법을 제안한다. 스트리밍 데이터의 컨텍스트(context)를 기반으로 개념 계층을 구성해서 사용자 개개인의 요청에 맞게 데이터를 전송, 그리고 동향이나 근사치 요청 질의에 효과적으로 대응할 수 있다.

2. 관련연구

기존에 연구되었던 스트리밍 데이터에 대해서 질의 처리를 하는 시스템들은 스트리밍 형태의 네트웍 데이터에 대한 질의 처리를 목적으로 개발된 STREAM시스템[1]과 센서 네트웍 환경에서의 스트리밍 데이터에 대한 질의 처리를 목적으로 개발된 Cougar시스템과Telegraph시스템[2] 그리고 스트리밍처럼 연속적으로 변화하고 무한한 인터넷과 웹(web)데이터의 질의 처리를 목적으로 개발된 Niagara시스템[3], 마지막으로 XML 스트리밍 데이터에 대한 질의 처리를 위한 시스템인 Tukwila시스템[4]등이 대표적이다. [표1]은 관련 시스템의 비교 분석한 내용이다.

1) 본 논문은 연구과제 R01-2003-1-000-10395-0인 한국과학재단의 위탁과제로 수행되었다.

[표 1] 스트리밍 데이터 처리 시스템의 비교

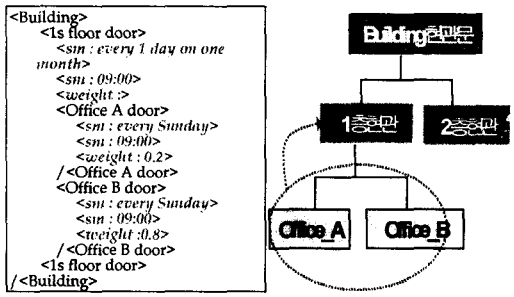
	XML-QL	SQL like	Hybrid	CQL
	Support	Not support	Support	Not support
	•source를 지정불가 •DTD만을 Graphic Interface에서 선택	•Sensor Network의 확장 •Eddy와 River를 이용 Source의 routing를 효과적으로 수행	•Focus on Streaming XML data •X-scan을 이용 tuple binding	•Network data를 대상 •기존의 DBMS의 기능을 사용하기 위한 DSMS의 형태
	Not support / Support	Support / Support	Support / Support	Support / Support
	Web data를 index형태로만 저장 (질의처리 불가능)	매우 제한적인 과거 데이터의 질의처리만 가능	Past data처리는 고려하지 않고 있다.	Analysis, Mining만을 위한 Disk로의 저장
	Impossible /Impossible	Possible (very restricted) /Impossible	Impossible /Impossible	Possible (Only mining) /Impossible
	NiagaraCQ	Eddy , River	X-Scan	DSMS

서로 다른 입력속도를 갖는 여러 개의 네트웍 데이터에 대해서 지연 없이 실시간 질의처리를 목적으로 개발된 STREAM와 Telegraph 시스템은 디스크와 같은 저장공간을 이용해서 데이터를 저장하고 저장된 데이터에 대한 질의를 허용하지만, 근래의 데이터만을 저장하고 저장된 데이터에 대한 질의역시 마이닝(mining)질의 만을 허용하고 있다. 그리고 스트리밍 데이터 중에서 스트리밍 XML 데이터만을 처리하기 위해 개발된 Tukwila 시스템은 실시간 질의와 연속적인 질의만을 고려하고 있기 때문에 과거 데이터에 대한 질의처리는 불가능하다는 한계가 있다.

3. 개념 계층을 이용한 스트리밍 데이터의 저장 및 관리

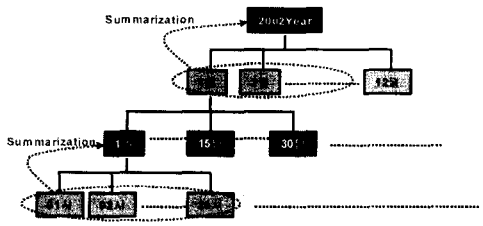
센서 네트웍, EDI(Electronic Data Interchange), LBS(Location-based services), Smart Building등 스트리밍 데이터가 사용되는 환경은 각 환경마다 정해진 형태 또는 DTD를 갖는 데이터를 사용하기 때문에 사용자가 환경에 맞게 개념 계층을 구성할 수 있다.

예를 들어 아래 [그림2]와 같이 건물의 현관문, 각 층의 현관문, 그리고 각 사무실의 출입문에 출입기록을 체크하는 센서가 설치되어서 출입기록을 저장하는 환경이라면 사용자는 XML형태로 개념 계층을 구성할 수 있다. XML형태로 표현된 개념 계층에서 가장 하위 레벨인 Office_A와 Office_B는 매주 일요일 9시에 상위 레벨인 1층 현관의 출입 정보로 요약이 된다(<sm : every Sunday> , <sm : 09:00>). 요약 시에 사용자가 Office_A의 출입 정보를 20%(<weight : 0.2>), Office_B의 정보를 80%의(<weight : 0.8>) 가중치로 요약을 수행한다. 그리고 요약이 완료된 후에는 요약된 정보에 해당하는 하위 레벨의 데이터는 삭제한다. 요약 후에 삭제된 데이터에 대한 질의는 요약된 정보인 1층 현관의 출입 정보로 대체된다. 마찬가지로 1층과 2층의 출입 정보도 매달 1일 오전 9시에 건물 현관문의 출입 정보로 대체된다.



[그림 2] 개념 계층의 XML형태의 표현의 예제

또한 동향이나 근사치를 요청하는 질의와 같이 시간을 기준으로 질의를 하는 경우가 많은 환경이라면 [그림3]과 같이 시간을 기준으로 저장되는 데이터를 시간의 개념 계층을 기준으로 요약이 가능하다. 또한 하루 단위의 날(day)에 대해서 주중(weekday)의 정보가 주말(weekend)의 정보보다 더 중요한 환경이라면 요약 시에 주중의 가중치를 80%로 정의하고 상대적으로 덜 중요한 주말의 가중치로 20%의를 정의해서 요약 시에 반영하는 것과 같이 사용자가 환경에 맞게 사용자 지정 속성을 적용할 수 있는 장점이 있다.



[그림 3] 시간을 기준으로 하는 개념 계층

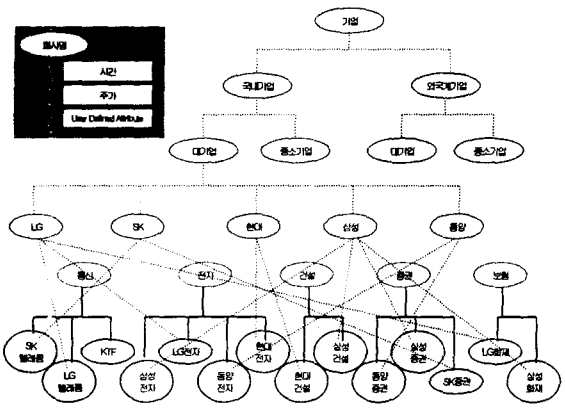
이와 같이 개념 계층을 이용한 스트리밍 데이터의 저장 및 관리는 스트리밍 데이터가 사용되는 모든 환경에 유연하게 적용할 수 있는 장점이 있는 반면 개념 계층을 사용자가 환경에 맞게 직접 설계를 해야 하는 문제점이 있다.

4. 개념 계층을 이용한 주식거래 시스템

스트리밍 데이터가 사용되는 환경으로 주식시장을 예로 들 수 있을 것이다. 주식시장에서 발생할 수 있는 과거 동향이나 과거의 시세에 대한 질의를 처리하기 위해서는 실시간 주식시세뿐만 아니라 과거 데이터에 대한 정보를 저장하고 있어야만 모든 질의처리가 가능하게 된다. 주식시세를 위한 개념 계층을 [그림4]와 같이 정의를 할 수 있다.

저장되는 데이터는 회사명, 시간, 주가 그리고 사용자 지정 속성을 저장한다. 최하위 레벨의 정보는 회사 각각의 주식 시세를 저장한다. 그리고 그 상위 레벨은 전자, 건설, 보험, 통신 등의 각 업종별로 요약을 통해서 해당 회사의 주식시세 정보를 저장한다. 그리고 다른 관점에서의 상위 레벨인 대기업의 그룹별로 요약을 통해서 주식시세 정보를 저장한다. 또한 사용자가 필요할 경우에는 중소기업, 그리고 국내 기업과 외국계 기업이라는 분류의 계층 구조마다

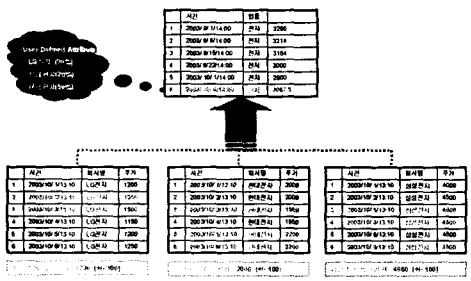
주식시세 정보를 저장할 수 있다.



[그림 4] 주식시장의 개념 계층

개념 계층을 이용한 기법의 가장 큰 장점은 사용자가 개념 계층을 직접 수정, 작성함으로써 모든 환경에 유연하게 대처할 수 있다는 점이다.

예를 들어 Kim이라는 사용자는 LG전자, 현대전자, 삼성전자의 주식시세가 30%, 20%, 50%의 비율로 요약된 주식시세의 일주일 단위의 동향을 분석하는 질의를 원하고 있고, 사용자가 원하는 형태의 요약(전주의 평균 시세보다 +/- 100 원 이상 차이가 나는 데이터만을 요약)을 원한다면 [그림5]와 같이 요약할 수 있을 것이다.



[그림 5] 사용자 지정 형태의 요약

사용자는 자신이 원하는 형태의 요약된 정보를 통해서 동향을 분석하고 대략적인 주가 예측이 가능하게 된다. "현재 전자 회사들의 주식 동향은 9월초부터 10월 첫째 주까지는 하락, 둘째 주부터 반등 중" 이라는 동향 분석이 가능하게 된다.

개념 계층을 이용한 증권시세 정보 시스템의 장점은 실시간 질의처리 뿐만 아니라 과거 데이터에 대한 질의처리가 가능하며 사용자의 특성을 반영할 수 있다는 장점이 있다. (현재 증권 시스템에서 사용되는 질의처리는 시스템에서 제공하는 질의에 대해서만 지원을 하지만 제안하는 기법을 사용한다면 사용자가 원하는 어떠한 질의의 형태라도 지원이 가능하다.) 또한 요약을 통해 오래된(long term) 데이터에 대한 근사치 요청 질의와 동향요청 질의가 가능하며 복잡한 질의의 형태에 대해서 Back ground processing을 통한 요약을 수행하기 때문에 질의에 대한 응답 시간이 향상된다.

이에 반해 한계점으로는 제안하는 기법을 효과적으로 사용할 수 있는 환경은 숫자로 표현되는 정보를 사용하는 환경으로 제한적이고 (변화하는 문자 정보를 저장하는 환경이라면 히스토그램과 같이 데이터의 분포도와 중복되는 데이터를 삭제하는 정도의 요약이 이루어지는 한계점이 발생) 문자열 인코딩 기법에 대한 추가적인 연구가 필요하다. 또한 데이터의 편차가 심한 환경이라면 요약된 정보에 대한 질의의 정확도가 떨어질 수 있다는 한계점이 있다.

5. 결론

과거 데이터에 대한 질의 처리를 위해서는 과거 RDB와 같이 디스크에 모든 데이터를 저장하는 것은 디스크의 물리적인 한계로 인해서 불가능하다. 그렇기 때문에 일정 시간마다 저장된 데이터를 어떠한 형태로든지 요약을 수행해야 한다. 이때 각 환경마다 사용되는 질의들이 다르고 사용자마다 원하는 데이터들이 다른 상황을 반영 해야만 한다. 이러한 문제들을 개념 계층을 사용자가 환경에 맞게 XML 형태로 작성함으로써 해결한다. 또한 사용자가 지정한 시간마다 정의된 형태로 요약을 수행하기 때문에 log(n)의 형태로 저장 공간의 낭비를 줄이고 많은 수의 데이터를 작은 수의 요약된 형태로 저장하기 때문에 질의에 대한 응답 시간도 log(n)의 비율로 감소하게 된다. 데이터의 분포도 정보를 저장하는 히스토그램과 비교하면 정보를 저장하는데 히스토그램보다는 많은 저장 공간을 사용하지만 시간을 중심으로 질의를 할 수 있다는 장점이 있다. 스트리밍 데이터의 대부분이 시간 정보를 포함하는 환경이기 때문에 시간을 기준으로 하는 질의처리는 사용자에게 질의처리에 있어서 많은 편리함을 제공한다. 또한 완벽한 정확도를 보장하지는 않지만 동향이나 근사치를 요청하는 질의와 같이 어느 정도의 오차를 허용하는 질의나 여러 데이터를 혼합해서 사용하는 복잡한 질의를 사용하는 환경이라면 개념 계층을 작성할 때 복잡한 형태의 데이터를 사용자 지정 속성을 이용해서 Back ground processing 형태의 요약을 수행하기 때문에 질의 시 응답 시간을 효과적으로 단축한다.

5. 참고문헌

[1] A. Arasu, B. Babcock. S. Babu, M. datar, K. Ito, I. Nishizawa, J. Rosenstein, and J. Widom, "STREAM: The Stanford Stream data Manager" , *IEEE Data Engineering Bulletin, Vol. 26 No. 1,* March 2003.

[2] Sirish Chandrasekaran and Michael J. Franklin, "Streaming Queries over streaming data" ,*VLDB*, Vol.12, Issue.2, 2002, pp.140-156.

[3] Jianjun Chen, David J. DeWitt, Feng Tian, Yuan Wang, "NiagaraCQ: A Scalable Continuous Query System for Internet databases" , *SIGMOD,* 2000, pp. 379-390.

[4] Zachary G. Ives, Alon Y. Halevy, Daniel S. Weld, "An XML Query Engine for Network-Bound data" , *VLDB,* Vol.11, Issue4., 2002, pp.380-40